Cancer InFocus: Tools for Cancer Center Catchment Area Geographic Data Collection and Visualization

Justin Todd Burus[1], Lee Park[2], Caree R. McAfee[1], Natalie P. Wilhite[1], and Pamela C. Hull[1,3]

ABSTRACT

Background: The NCI added Community Outreach and Engagement (COE) requirements for NCI-designated cancer centers in 2017, including the charge of characterizing the cancer burden in the geographic area served by their center (i.e., catchment area). Doing so helps cancer centers better identify needs and areas of inequality in their populations to guide research and outreach. To accomplish this, current and comprehensive data must be gathered from multiple sources and analyzed by the COE—a task that is tedious and inefficient. In this paper we present an efficient solution, known as Cancer InFocus, to collecting and visualizing quantitative data that we have generalized for use by other cancer centers on their catchment areas.

Methods: Cancer InFocus utilizes open source programming languages and modern data collection techniques to gather and transform publicly-available data from various sources for use in specific geographic contexts.

Results: Cancer InFocus delivers a choice of two routes for creating interactive online mapping applications that visualize cancer incidence and mortality rates, along with relevant social determinant and risk factor variables, at various geographic levels for a defined cancer center catchment area.

Conclusions: Generalized software has been produced to collect and visualize data on any set of U.S. counties, which can be automated to continue providing the most up-to-date data.

Impact: Cancer InFocus provides tools for cancer centers to accomplish the critical task of maintaining current and comprehensive catchment area data. The open source format will facilitate future enhancements through user collaboration.

Introduction

Starting in 2012, the NCI laid out an expectation for each of their comprehensive and clinical cancer centers to establish a catchment area— a "self-defined geographic area that the Center serves or intends to serve in the research it conducts" (1). Engagement with the catchment area was further reinforced with the creation of a Community Outreach and Engagement (COE) component to the NCI's Cancer Center Support Grant requirements in 2017, which includes the expectation to "analyze the demographics and cancer burden of (the cancer center's) catchment area" (1). This enables cancer centers to grasp not only the overall cancer burden and disparities in their communities but also to recognize barriers to cancer prevention, control and care that exist. Developing the ways in which we engage with our community to successfully address unequal outcomes through outreach and research requires understanding the multitude of factors that are influencing them (2). To come to such an understanding, it is necessary for cancer centers to have current and complete knowledge of their catchment area, and thus current and complete data (3). Thankfully, much of the relevant quantitative data are publicly available, and housed in just a few key sources. Acquiring and preparing this data for a specific catchment area, however, tends to be a tedious and inefficient process—a process which gets repeated across the 63 NCI-designated cancer centers with identified catchment areas, utilizing the same sources but approaching them with different methodologies and geographic areas to consider.

When the University of Kentucky (UK; Lexington, KY) Markey Cancer Center's Community Impact Office began gathering data for assessing the cancer burden in our own catchment area in early 2021, we sought to develop a more efficient way of completing this task. Many of the quantitative variables we wanted to include in our catchment area assessment could be accessed in publicly-available and machine-readable formats thanks to the OPEN Government Data Act of 2018 (4). This means that, rather than collecting data by manually visiting a website, downloading datasets onto a computer, and processing them with statistical software, we could instead write computer programs to utilize certain tools and techniques for both fetching all of the data and outputting it in our desired format. This would give us a reusable software solution to the problem which circumvents the need to start over each time a data update comes available. Furthermore, this would facilitate moving the data from spreadsheets into a visual interface by combining the dataset creation programs with available geographic information system (GIS) mapping tools. Cancer center staff, researchers, community organizations and even lay persons could use this interface to explore and generate custom views of the area's cancer burden and relevant influencing factors. With all this in mind, our staff set out to create such a solution, with the added intention of designing it so that generalized versions could be easily disseminated to and adopted by other cancer centers for their own catchment areas.

The end result of these efforts is a software collection known as Cancer InFocus. Cancer InFocus automates the location, download, manipulation, and visualization of publicly-available data pertaining to cancer incidence and mortality, population demographics, social determinants of health, relevant behavioral risk factors, and healthcare resources, at various geographic levels, across a set of U.S. counties. This process performs the initial data capture and visualization, and can be set up to make regular, automatic updates to each with minimal

[1]Markey Cancer Center, University of Kentucky, Lexington, Kentucky. [2]Department of Statistics, University of Kentucky, Lexington, Kentucky. [3]Department of Behavioral Science, College of Medicine, University of Kentucky, Lexington, Kentucky.

Corresponding Author: Justin Todd Burus, University of Kentucky Markey Cancer Center, 760 Press Avenue, Suite 460, Lexington, KY 40536. Phone: 859-562-0291; E-mail: tburus@uky.edu

Cancer Epidemiol Biomarkers Prev 2023;32:889–93

doi: 10.1158/1055-9965.EPI-22-1319

Table 1. Publicly available data sources considered for use in Cancer InFocus.

Source	Agency	Domain(s)
American Community Survey	United States Census Bureau	Sociodemographics, Economics & insurance, Housing & transportation
BRFSS	CDC	n/a[a]
Certified Mammography Facilities	FDA	Locations
Food Atlas	United States Department of Agriculture Economic Research Service	Environment, Geographies
Form 477 Broadband Availability	Federal Communication Commission	n/a[b]
Health Center Service Delivery Sites	Health Resources and Services Administration (HRSA)	Locations
Health Professional Shortage Areas	HRSA	Locations
Labor Force	Bureau of Labor Statistics	Economics & Insurance
Lung Cancer Screening Registry	American College of Radiology	Locations
National Plan & Provider Enumeration System	Health & Human Services	Locations
Places	CDC	Screening & Risk Factors
Safe Drinking Water Information System	Environmental Protection Agency	Environment
State Cancer Profiles	NCI, CDC	Cancer Incidence, Cancer Mortality
Wonder	CDC	n/a[a]

[a]Source not used by Cancer InFocus because variables of interest were not available below the state level.
[b]Source used by Cancer InFocus, but not included in interactive application.

effort. The objectives of this article are (i) to describe the methodology that the UK Markey Cancer Center's Community Impact Office used to create Cancer InFocus as an efficient solution for data capture and visualization, and (ii) to outline the steps that other cancer centers can take to apply Cancer InFocus to their respective catchment areas. Cancer InFocus results in a tremendous time-savings in catchment area characterization for cancer centers as they strive to serve their communities in a fight against the second-leading cause of death in America.

Materials and Methods

Identifying sources

The first challenge in this project was to define the desired quantitative data, locate data sources, and determine whether an automated means of retrieving these sources was available. These efforts helped us identify several prospective data sources for inclusion in Cancer InFocus (**Table 1**). All of these sources permitted some form of automated data collection through application programming interface (API) calls, URL queries, HTTP downloads and/or web scraping, though we chose to exclude a few because the variables of interest in them were not publicly available below the state level (CDC Wonder, CDC BRFSS).

Programming language and software solutions

The scripts to execute data collection and manipulation were written in the Python programming language (version 3.9, RRID: SCR_008394)—an object-oriented programming language that is widely used by both data scientists and computer developers. This choice provided the flexibility to perform API calls, automate web scraping within a Google Chrome browser and manipulate datasets, all with relative ease. Python does require more advanced programming skills than is likely available within most COE offices, and yet is well-known enough that we would expect any NCI-designated cancer center to have staff who can competently execute these files. We also developed a version of the programs using Google Colab, which allows individuals to run Python in a cloud environment without local installations of the software or knowledge of the language itself (though personally troubleshooting any issues would be a problem in this circumstance).

Ways to visualize the data in their geographic context were then built using two different platforms: ArcGIS Online and R's Shiny (R Project for Statistical Computing, version 4.2.1, RRID:SCR_001905; Shiny version 1.7.2, RRID:SCR_001626). ArcGIS Online is a subscription-based GIS mapping service utilizing cloud storage that features robust Python integration and a JavaScript API for creating customized online mapping applications. Shiny is an R package which allows for the creation of interactive web applications using the R programming language. Although ArcGIS Online provides a more feature-rich system for working with the data both in and beyond the scope of this project, having an open source mapping solution in Shiny is desirable to reduce barriers to wider adoption of this process by other cancer centers.

Additional notes

During the course of this project we produced comma-separated values (CSV) files for the catchment areas of all 63 comprehensive or clinical cancer centers as enumerated by DelNero and colleagues (5). We also produced geographic area files for the entire United States and the Appalachian region as defined by the Appalachian Regional Commission (6). Cancer InFocus is available through a no cost licensing agreement with the University of Kentucky. Interested users can contact the authors to request access.

Data availability

No datasets were generated or analyzed during this study.

Results

The above efforts resulted in two different three-stage sequences of tools to perform data collection, manipulation, export, visualization, and application (**Fig. 1**). We will first focus on the route that utilizes ArcGIS Online. The three stages involved in this process are as follows: (1) two Python programs used in tandem to gather and

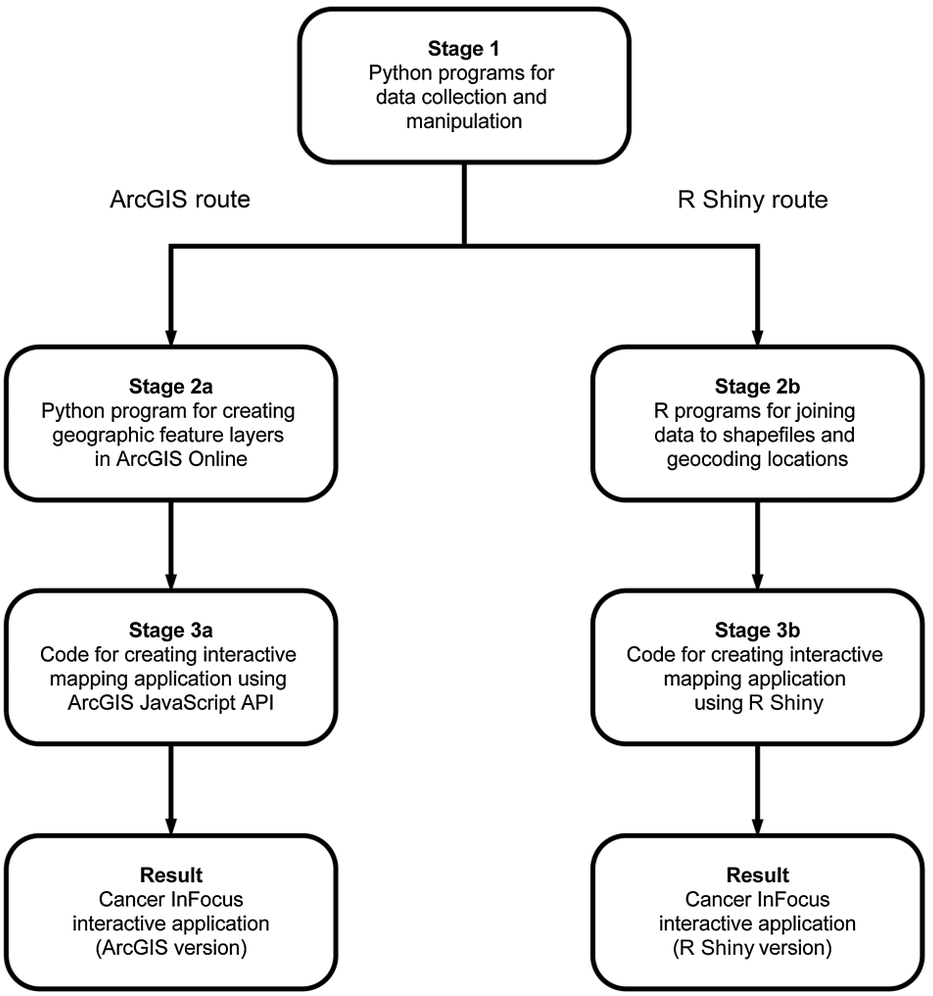

Figure 1.
Cancer InFocus workflow diagram. Stages in the process of developing the final Cancer InFocus application using either ArcGIS or R Shiny are detailed, including relevant programming languages employed.

manipulate catchment area data, (2a) a Python program for creating geographic feature layers in ArcGIS Online, and (3a) a set of custom HTML code for creating an online mapping application. Users should start at Stage 1 of the tools, but can terminate their use at any step along the way as fits their needs.

The first program in Stage 1 is a technical file defining various classes, functions, and methods to perform generalized gathering and manipulation of the data. This script gets loaded as a module into the second program of Stage 1, alongside user input regarding their catchment area. Together they define the specific geographic boundary for data collection and produce the final output. Runtime on this stage generally takes between 5 to 20 minutes (depending on size of the defined geographic area, speed of the user's connection and the processing capabilities of their machine) and results in 26 CSV files, broken down by geographic levels and broad categories of variables, and displayed in both wide and long data formats. Alternatively, the user can adjust the code to call a function which outputs two Excel workbooks containing the formatted and categorized values across multiple sheets. When tested on Markey Cancer Center's catchment area—the state of Kentucky—Stage 1 captured 110 different variables across four geographic levels (county, Census tract, Census block, and location), resulting in 26 CSV output files and approximately 367,000 unique values, averaging a runtime of seven minutes and 18 seconds across five iterations.

Stage 2a involves running a program to attach the output from above to shapefiles of the desired geographic area, along with a selection of colors and/or symbols, to produce several web maps on the ArcGIS Online server. Runtime on map creation varies depending on size and complexity of the geographic area and speed of communication with the online servers. The maps created in Stage 2a exist permanently within ArcGIS Online, allowing them to be incorporated into other projects and analyses beyond the interest of this paper.

In Stage 3a, custom HTML code accesses the maps from Stage 2a and creates an interactive JavaScript interface for viewing them online. Users of the resulting application also have the option to capture and save digital images of the maps and download the datasets underneath the maps being shown. Implementation of this code requires users to adjust certain portions to fit the specifics of their geographic area and use case.

In a normal, first-time workflow progressing from Stages 1 through 3a, datasets are created, followed by maps of the variables gathered, and then a JavaScript application ready to be embedded in a website. Should a user want to update data at a later time, Stages 1 and 2a can be run again. Stage 2a will overwrite the layers from previous runs, which propagates updates to the application generated in Stage 3a. This

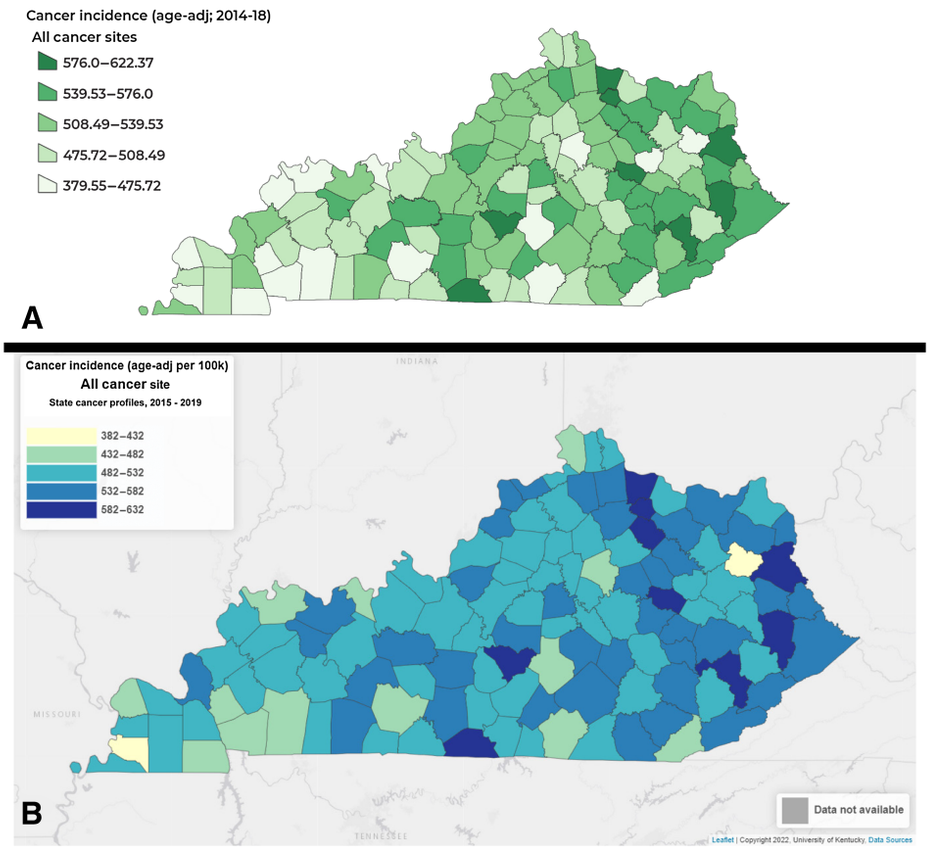

Figure 2.
Examples of output from Cancer InFocus applications. **A,** Output generated by ArcGIS JavaScript application. **B,** Output generated by R Shiny application.

allows users to quickly bring in the latest data and schedule automated updates to their online application, if desired.

The alternative option of using Shiny incorporates Stage 1 from the above, followed by: (2b) two R programs which bind together data with shapefiles by geographic level and geolocates facilities and providers, and (3b) a Shiny program which defines and launches an interactive Shiny mapping application. Shiny web applications can also be set to update as above by automating Stages 1 and 2b and then pushing the new data into the application's www/ folder.

Examples of output from the final interface for both ArcGIS Online and Shiny versions of Cancer InFocus: Kentucky are given in **Fig. 2**. The included data sources are listed in **Table 1**, along with the domain where they can be found within the application. A live version of the Shiny application can be viewed at https://cancerinfocus.uky.edu/.

Discussion

As stated above, having accurate and timely data is crucial for a cancer center wanting to better engage with the cancer-related needs of the population it is serving. Previous studies have used publicly available data to better understand and address specific aspects of the cancer burden in a geographic area, such as environmental exposures (7, 8), social risks (9), access to and use of technology (10), and screening behaviors (11). Cancer InFocus captures versions of all of these factors and more for the given catchment area input.

Since releasing Cancer InFocus: Kentucky in the summer of 2022, the UK Markey Cancer Center's Community Impact Office has used it as a key strategy for disseminating catchment area data internally to the cancer center's leadership, researchers, staff, and trainees to guide development of research questions and strategic plans that respond to catchment area needs. The application has provided a self-service way for these parties to obtain relevant catchment area data and data visualizations, reducing the need for our team to fill individual requests for assistance with descriptive data and maps for grant applications, presentations, and other uses. Our team has also used the application within our cancer center outreach programs to gain insights for better targeting interventions to specific populations and communities with greatest need. Moreover, we have disseminated Cancer InFocus: Kentucky externally to community partner organizations across the state for a variety of uses, such as program planning, grant applications, and community health needs assessments led by public health departments and hospitals.

Modern advances in computational and statistical software have made obtaining catchment area data much simpler than it was for previous generations, but certain inefficiencies have persisted. The tools developed here demonstrate a marked improvement on the status quo, allowing on-demand data collection for many of the relevant quantitative variables that a cancer center might want to include in their catchment area assessments. The resulting outputs can be easily transformed into deliverables for broader communication thanks to their organization by geographical identifiers at the county, Census tract and Census block group levels. Such organization facilitates the ability to view emerging spatial patterns in the data outside of what may be picked up by numerical analysis alone. These tools not only save time but fundamentally level the playing field between cancer centers who have plentiful resources available for data collection and analysis and those who do not (12). Moreover, adaptations of these tools could even prove beneficial for population-level research in other disease fields

(such as cardiovascular disease) due to large overlaps with the relevant social determinants of health and risk factors related to cancer.

Cancer centers can decide to use either ArcGIS Online or Shiny for producing the final interactive mapping application, weighing the pros and cons of each. We initially used ArcGIS Online for completing this task based off of familiarity with the tools and the depth of what can be done on that platform. This also worked well with Stage 1 of the process, as both the data collection and map creation could be completed using only the Python programming language. Moreover, manual additions could be made to the application by building maps in ArcGIS Pro (ESRI's desktop mapping software), uploading them into ArcGIS Online and referencing them from the JavaScript code. The layers produced for the ArcGIS web application exist permanently in the ArcGIS Online cloud environment, meaning that they can be accessed for additional projects and analyses apart from the application itself. Some limitations of this option, however, are that designing the final public-facing application requires coding a unique JavaScript webpage from scratch, and the fact that subscriptions to ArcGIS Online are expensive and possibly not available to some potential users. This last consideration is what led us to develop a Shiny version as well. Building the final product in Shiny is much simpler and faster than creating multiple permanent map layers and crafting a whole JavaScript webpage from scratch to display them, and costs nothing to complete in its entirety (though paid options for deploying Shiny applications exist). The disadvantages of utilizing Shiny are that the map layers produced only exist during the execution of the program itself, that this route requires changing coding languages during implementation, and that large-scale public distribution may be difficult to accomplish without leveling up into one of the paid deployment tiers.

The tools described in this paper have important limitations to consider. First, the nature of the solutions provided requires a nontrivial amount of technical knowledge on behalf of the user, though some of this need has been obviated by leveraging public cloud-computing resources. Additional development could be done to include these tools in a graphical user interface that would make the process manageable for most any cancer center staff. That said, the combination of an ever-evolving digital data landscape and the growing mission of cancer center offices of Community Outreach and Engagement makes it ideal for these offices to consider including a staff person dedicated to managing catchment area data collection and visualization tools. Second, some datasets accessed by these tools may undergo revisions at future points that render the current methods of obtaining them nonoperational. The desire of our team is that, by making the code contained within these tools public, we will be able to build a user community that can maintain them and adapt with any changes that arise. Third, some desirable sources (such as complete state cancer registry data) are only accessible through individual data requests, and as such would need to be added into the datasets by end users who want to include them. Providing the outputs in simple long and wide formats was done in order to make linking with individually accessed data easier. Lastly, we are unable to guarantee the accuracy of data obtained by these tools, but rather cite the sources accessed as the final authority on their veracity.

Cancer is a deadly disease that requires significant resources to address. Cancer InFocus creates an easier way to access quantitative data and put it to use in identifying where opportunities, needs and disparities exist, allowing more time to be spent gathering supplemental qualitative data and engaging with the community in the ongoing effort to reduce this burden.

Authors' Disclosures

J.T. Burus reports grants from NCI during the conduct of the study. L. Park reports grants from NCI during the conduct of the study. C.R. McAfee reports grants from the NCI during the conduct of the study. N.P. Wilhite reports grants from NCI during the conduct of the study. P.C. Hull reports grants from NCI during the conduct of the study.

Authors' Contributions

J.T. Burus: Conceptualization, software, visualization, methodology, writing–original draft, project administration, writing–review and editing. **L. Park:** Conceptualization, software, visualization, methodology, writing–original draft, writing–review and editing. **C.R. McAfee:** Conceptualization, writing–review and editing. **N.P. Wilhite:** Conceptualization, writing–review and editing. **P.C. Hull:** Conceptualization, supervision, writing–original draft, writing–review and editing.

Acknowledgments

P.C. Hull and C.R. McAfee were supported by the NCI of the NIH grant P30CA177558. P.C. Hull also received support from the William Stamps Farish Endowed Chair in Cancer Research. This research was also supported by the Cancer Research Informatics Shared Resource Facility of the UK Markey Cancer Center (P30CA177558).

Received December 16, 2022; revised March 21, 2023; accepted May 8, 2023; published first May 10, 2023.

References

1. NIH. NIH guide for grants and contracts; 2022. Available from: https://grants.nih.gov/grants/guide/pa-files/PAR-21–321.html.
2. Doykos PM, Chen MS Jr, Watson K, Henderson V, Baskin ML, Downer S, et al. Recommendations from a dialogue on evolving National Cancer Institute-Designated Comprehensive Cancer Center community outreach and engagement requirements: a path forward. Health Equity 2021;5:76–83.
3. Tai CG, Hiatt RA. The population burden of cancer: research driven by the catchment area of a cancer center. Epidemiol Rev 2017;39:108–22.
4. H.R.1770 - 115th Congress (2017–2018): OPEN Government Data Act. (2017, March 29). Available from: https://www.congress.gov/bill/115th-congress/house-bill/1770.
5. DelNero PF, Buller ID, Jones RR, Tatalovich Z, Vanderpool RC, Ciolino HP, et al. A national map of NCI-Designated Cancer Center catchment areas on the 50th anniversary of the cancer centers program. Cancer Epidemiol Biomarkers Prev 2022;31:965–71.
6. Appalachian Counties Served by ARC - Appalachian Regional Commission. Available from: https://www.arc.gov/appalachian-counties-served-by-arc/.
7. Ruano-Ravina A, Aragones N, Kelsey KT, Perez-Rios M, Pineiro-Lamas M, Lopez-Abente G, et al. Residential radon exposure and brain cancer: an ecological study in a radon prone area (Galicia, Spain). Sci Rep 2017;7:3595.
8. Joseph N, Kolok AS. Assessment of pediatric cancer and its relationship to environmental contaminants: an ecological study in Idaho. Geohealth 2022;6:e2021GH000548.
9. Beyer KMM, Kasasa S, Anguzu R, Nambooze S, Amulen PM, Jankowski C, et al. Parish level social factors predict population-based cervical cancer incidence in Kampala, Uganda, 2008–15: an ecological study. Lancet Glob Health 2022;10:S12.
10. Vanderpool RC, Stradtman LR, Gaysynsky A, Chen Q, Johnson M, Huang B. Access to and use of technology for health: comparisons between Appalachian Kentuckians and the general U.S. population. J Appalach Health 2021;3:60–73.
11. Harper DM, Plegue M, Jimbo M, Gorin SS, Sen A, Schlecht N. US women screen at low rates for both cervical and colorectal cancers than a single cancer: a cross-sectional population-based observational study. Elife 2022;11:e76070.
12. Beyer K, Kasasa S, Anguzu R, Lukande R, Nambooze S, Amulen PM, et al. High-resolution disease maps for cancer control in low-resource settings: a spatial analysis of cervical cancer incidence in Kampala, Uganda. J Glob Health 2022;12:04032.